Data mining in Bioinformatics

Viktória Bendíková, Andrej Černek, Ondřej Lošťák, Dávid Meluš and Šimon Varga

Faculty of Informatics Masaryk University

Introduction

In recent years, the amount of biological data has grown exponentially, which raises a major challenge: How to efficiently extract useful and biologically meaningful information from these available data? The field of machine learning significantly contributes to the development of approaches to transform the data into biological knowledge about the underlying mechanism.

We provide a brief overview of various machine learning techniques focusing mainly on their application to different types of bioinformatics data from major biological domains.

DNA/RNA Sequences

Nucleic acid sequencing is the process of determining the order of nucleotides in the nucleic acid sequence. The declining cost and increased efficiency of sequencing results in a tremendous amount of sequential data. This data has to be analyzed which consists of the following problems:

- **Gene prediction**
  - identification of introns (non-coding regions) and exons (coding regions) in DNA
  - identification of active sites, gene structures, open reading frames and regulatory elements
  - identification of repeats
- **Identification of the biological function**
- **Comparison**: finding similarity between sequences, homology

The machine learning techniques are mainly used in gene prediction. The most common approaches are based on Hidden Markov models, conditional random fields, Support Vector Machines (SVMs), neural networks or their combination. In the case of homology, inductive logic programming can be used.

Protein Sequences

The primary structure of proteins is a chain of amino acids bonded via peptide bonds, which is represented by a sequence of letters over a 20-letter alphabet, corresponding to 20 naturally occurring amino acids. These sequences are retrieved through protein sequencing. Major problems solved based on protein sequences involve the prediction of crucial protein properties:

- **Protein Function Prediction** consists in assigning biological or biochemical roles to proteins.
- **Protein Structure Prediction** is a complex biological problem, which is approached at different levels: 1-D prediction of protein structural features, 2-D prediction of spatial relationships between amino acids and 3-D prediction of the protein tertiary structure.

The most successful machine learning techniques aiming at solving the mentioned problems from the proteomics domain involve Support Vector Machines, neural networks, and Hidden Markov models.

Microarray

A microarray is a method used for observing the expression (activity) of many genes at the same time. In this method two samples (reference sample and experiment sample) are usually compared, so we typically collect mRNA molecules from both samples.

For every gene we know if it is expressed only in the reference sample, only in the experimental sample or both samples.

- **Gene Functional Class Prediction**
  Clustering methods help to construct co-expression networks. Clusters are created from genes expressed together and this can predict the function class of a gene.
- **Disease Characterization and Detection**
  For some diseases are typically expressed some specific genes. This can be used for supervised classification and detection of disease. On the other hand clustering can identify anomalies.

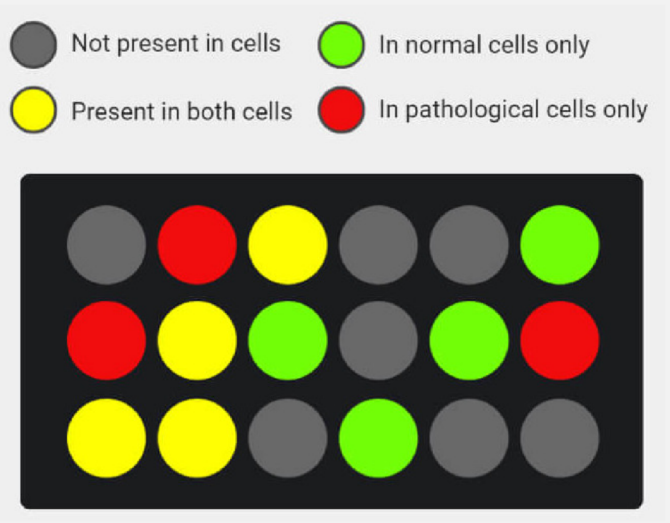

Figure: Microarray schema.

Microarray Time Series

A single gene expression matrix is like a cell's snapshot in one time point. Taking several snapshots over time may result in more detailed insight into the functioning of an underlying biological system.

- **Genetic regulatory network (GRN) inference**
  GRNs can be used, for example, to design experiments or drugs. If only one biological sample is of interest, the data is a matrix consisting of expression levels for every gene and time point. Here take place various clustering ML algorithms.
- More challenging is **triclustering**, needed for comparative analysis of tested biological samples (e.g. cells in different environments). Firstly, the data is of 3 dimensions: gene, sample, and time; thus having high complexity. Secondly, clustering reveals similar patterns, but finding differential expression patterns is desired.

Biomedical Image Analysis

Biomedical images correspond to a wider category of image types acquired from biological systems. Particularly of interest are so-called '*Medical Images*', related to the medical practice such as diagnosis, treatment, and follow-ups.

These images are acquired through mechanisms that range from simple digital cameras to complex medical systems like Computed Tomography Scanning, Ultrasound or Magnetic Resonance Imaging.

The basic applications of Machine Learning in this field are for example:

- **Computer assisted diagnosis**: These kinds of systems provide a high-level concept identification in the images that may support the decision-making process of the specialists.
- **Automatic quantification** is the process of identifying zones of interest by dividing the image into disjoint zones according to the extracted features like colour or texture attributes.

Due to high dimensionality of the image data types and frequent non-linearity of the features of interest, the greatest results are yielded by kernel-based methods like SVMs and deep learning methods, particularly Convolutional Neural Networks.

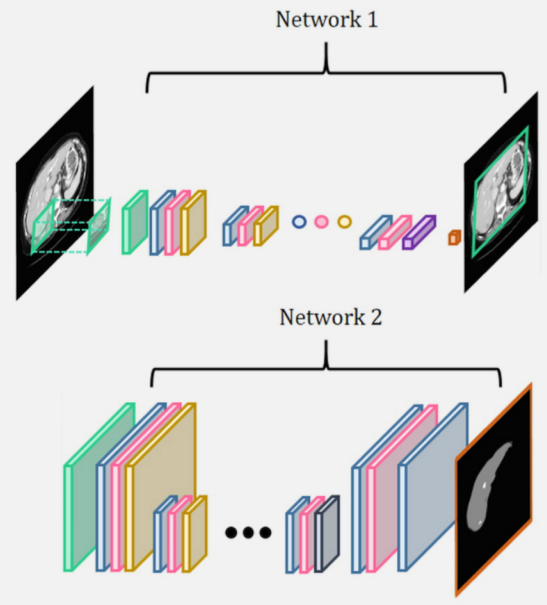

Figure: Cascaded CNN network for tumor segmentation.